ChinaXiv合作期刊

知识管理论坛 ISSN 2095-5472 CN11-6036/C
Knowledge Management Forum

E-mail: kmf@mail.las.ac.cn
http://www.kmf.ac.cn

【学术探索】

中文社交媒体用户性别预测研究

——以新浪微博短文本内容为例

刘雅琦[1] 李得志[2] 王瑞雪[3]

1. 中南财经政法大学信息与安全工程学院 武汉 430073
2. 百度网讯科技有限公司 北京 100085
3. 武汉大学信息管理学院 武汉 430072

摘要：[目的 / 意义] 与互联网的高速发展不同，个人信息安全保护的发展相对滞后，通过预测社交媒体用户的性别，能够更好地针对不同性别用户提供隐私保护。[方法 / 过程] 以新浪微博这一社交媒体中用户发布的短文本为研究对象，从中抽取语言特征和主题特征，为每一个用户构建基于语言特征、主题特征以及两个特征叠加的特征表达向量，利用 SVM 机器学习算法构建性别预测的分类器。[结果 / 结论] 实验表明，从微博短文本中抽取的语言特征和主题特征能够准确预测用户性别，其效果在主要评价指标中均有大幅提升。

关键词：短文本　性别预测　主题特征　语言特征

分类号：TP391.1

引用格式：刘雅琦 , 李得志 , 王瑞雪 . 中文社交媒体用户性别预测研究 : 以新浪微博短文本内容为例 [J/OL]. 知识管理论坛 , 2021, 6(4): 213-227[引用日期]. http://www.kmf.ac.cn/p/255/.

1 引言

随着互联网的深入发展，近年来信息安全逐步得到了人们的重视，中共中央成立了中央网络安全和信息化领导小组，“没有信息安全就没有国家安全”的理念深入人心。但现阶段，对信息商业价值的利用仍远远超过了对信息隐私安全的保护，信息的隐私保护依然处于相对滞后的状态；公共部门信息资源增值利用中，个人信息还存在着信息授权、利益平衡、法律救济和监管多方面的风险[1]。现有的法律体系中，虽然有大量的法律法规对个人信息保护提出立法，但在实际过程中，法律法规起到的保护作

基金项目：本文系国家社会科学青年基金资助项目“大数据环境下基于个体识别风险的个人信息利用研究”(项目编号：14CTQ016）研究成果之一。
作者简介：刘雅琦（ORCID：0000-0002-2361-2363），副教授，博士；李得志（ORCID：0000-0002-1754-1474），硕士研究生；王瑞雪（ORCID：0000-0001-5932-9036），博士研究生，通讯作者，E-mail：ruixue_wang@whu.edu.cn。
收稿日期：2021-07-05　　发表日期：2021-08-24　　本文责任编辑：易飞

用有限，个人信息的保护还存在一些障碍 [2]。

社交媒体持续发展，用户数量不断壮大。一方面社交媒体的发展为用户提供了方便快捷的信息获取方式；另一方面由于社交媒体的使用者门槛较低，社交网络的开放性、共享性与连通性的特点 [3]，使得用户的个人信息容易受到侵犯。为保护个人信息安全，部分用户在进行注册时会选择不填或虚假填写自己的性别 [4]，而相关研究表明女性用户对信息层面因素敏感，更易受影响 [5]，相较而言更容易透露自己的隐私信息 [6]。因此需要基于用户的性别提供服务，对用户进行适当的信息保护，使用户免受互联网中大量垃圾信息的伤害，如不对女性群体进行暴力内容的推送等。与此同时，用户的性别信息也是用户画像的重要组成部分，准确的用户画像可以为企业营销、广告投放、内容推荐提供便利 [7]；用户也可以从中获得个性化推荐内容，减少信息搜寻的时间，提高使用社交媒体的满意度。

近年来，用户画像相关的测评比赛也广泛兴起，例如名为 PAN 的学者群体举办了 6 届作者特征提取测评和 1 届僵尸用户与用户性别测评 [8]，由中国中文信息学会社会媒体处理专委会主办的全国社会媒体处理大会（SMP）于 2016-2018 年连续三年组织了相关的用户画像比赛 [9]。相关测评比赛中，性别预测是重要的子任务，是用户画像的核心内容之一，也是其他应用的基础 [10]。之所以要进行社交媒体用户的性别预测，是因为用户在进行注册时会忽略性别、兴趣等相关信息 [11-12]。

本文以新浪微博这一社交媒体平台中的用户信息为研究对象，利用不同性别用户语言表达和兴趣偏好上的差异预测用户性别。在社交网络中，男性和女性用户使用的语言以及兴趣爱好具有差异，A. H. Schwartz 等 [13] 从 75 000 名志愿者的 Facebook 消息中收集了 7 亿个单词、短语和主题实例，对其分析显示，不同性别的用户使用的语言有很大的不同，语言和性别以及年龄之间有着比较大的关联；M. Vicente 等 [14] 对 65 000 名英语用户的用户名、用户描述、图片和发送的推特内容进行分析，发现性别对用户的语言使用有影响，从而实现对用户性别的预测。因此，用户发布的社交媒体内容与性别相关联，呈现出差异化的特点。在此基础上，本文通过分析不同性别用户在发送社交媒体短文本时的差异，提取相关的语言特征和主题特征，构建模型进行用户性别的预测。

2 相关研究

2.1 基于图像的性别预测

基于图像的性别预测是通过分析用户的面部特征进行预测。目前，基于图像的用户性别预测主要使用的是传统图像分类方法，即通过模型提取图像中的人脸特征，再利用分类算法进行预测。常用于提取人脸特征的模型有 BIF（Bio-inspired Features）[15-16]、主动外观模型（Active Appearance Model, AAM）[17]、局部纹理特征（Local Binary Pattern, LBP）[18-19] 等。完成人脸特征提取后，利用不同的算法进行分类，常使用的算法有 k- 近邻 [18]、SVM 算法 [19]、AdaBoost 算法 [20] 等。近些年，随着深度学习在图像识别上的发展，各种神经网络算法 [21-22] 在基于图像的性别预测研究中取得了不错的效果。

2.2 基于用户信息的性别预测

在社交网络中，基于用户信息的性别预测主要分为两类，一类是基于用户的公开信息进行预测，另一类为基于用户发表的短文本内容进行预测。

2.2.1 基于用户公开信息的性别预测

基于用户公开信息的性别预测利用用户的账户名称、个人描述、个人主页设置、标签等信息，如 J. D. Burger 等 [23] 使用 Twitter 用户的账户名称、个人描述等用户公开信息预测用户的性别，最高可达 92% 的准确率；J. S. Alowibdi 等 [24] 提取了用户在 Twitter 上 5 个不同位置设置的颜色：个人资料背景颜色、文字颜色、链接颜色、边框填充颜色以及界面边框颜色做

知识管理论坛，2021（4）:213-227
DOI: 10.13266/j.issn.2095-5472.2021.021

为特征预测用户的性别，在不同数据集大小的实验中基本都能达到 70% 左右的准确率。社交媒体中存在大量缄默用户，其特点为很少发表内容、微博标签较少，因此准确预测较难，钱铁云等 [25] 利用微博用户个人资料中的标签信息，对缄默用户进行性别预测，达到了 71% 的准确率。

当用户的公开信息特征与训练样本的特征之间差异较大时，基于用户公开信息的性别预测方法的准确率会降低；同时用户公开信息量较少也会影响预测结果，例如用户昵称简短、没有个人描述等。此外，用户出于个人信息隐私保护的原因，在个人主页设置中选择不公开个人信息，将会使预测准确率大幅下降。

2.2.2 基于内容的性别预测

文本内容可根据长度不同分为短文本与长文本，社交媒体的文本主要为短文本，包括原创文本、转发文本以及评论文本三种类型。S. Li 等 [26] 提出了一种整数线性规划方法（Integer Linear Programming），利用用户原创及转发文本中的评论交互文本预测用户性别；戴斌等 [27] 利用半监督学习的方法实现了基于短文本内容的用户性别预测，达到了 84.3% 的准确率，解决了监督学习方法需要人工标注样本的障碍；N. Cheng 等 [28] 从 Twitter 文本中抽取了用户语言的心理语言学特征用于构建特征空间进行用户性别预测，达到了 85.13% 的准确率；J. A. B. L. Filho 等 [29] 把用户发送的 Twitter 文本中的字词个数、标点符号等作为文本元属性，进行用户性别预测，其准确率达到了 81.6%；Q. Wang 等 [30] 对比了文本表示方法 VSM（Vector space model）与主题模型 LDA（Latent Dirichlet allocation）、LSA（Latent semantic analysis）预测中文社交媒体中的用户性别、地域和年龄相关的人口统计学信息的效果，主题模型 LSA 在性别预测上效果表现最好，准确率达到 87.2%，但相较于 LDA 与 VSM 效果提升也比较有限。

n 元语法模型是自然语言处理中常用的模型，在性别预测领域有大量的研究以此为基础进行短文本分析，进而预测用户性别，例如 C. Peersman 等 [31] 使用 n 元语法模型并用卡方检验进行特征选择，利用构造的特征向量进行用户性别和年龄的预测；王晶晶等 [32] 在 n 元语法特征的基础上加上了首尾特征，使用用户的姓名和微博内容对性别进行预测，当用户样本足够大时，将基于用户姓名的分类器和基于微博内容的分类器融合之后能达到 90% 的准确率；Z. Miller 等 [33] 使用 n 元语法特征结合贝叶斯算法来预测用户的性别，其使用了 6 种特征选择方法，最高可以达到 97% 的准确率；D. Rao 等 [34] 抽取了用户的社会语言特征并与 n 元语法特征结合对 Twitter 用户的性别、年龄、地域和政治倾向进行了预测，对性别的预测准确率为 72%。

基于内容的性别预测方法对文本内容量的需求较高，社交网络中用户发送的文本多以短文本为主，当用户发送的内容较少时，仅凭借少量的文本内容很难准确预测用户的性别，这要求进行性别预测时所选取的文本特征既要体现出性别差异，也要有足够大的使用率。当数据量不足时会出现构建的分类器属性稀疏等问题，导致性别预测的准确率下降。

3 实验数据与预处理

本文使用中文社交媒体平台新浪微博的用户数据，数据集来源于“SMP CUP2016 微博用户画像”比赛 [35]。数据集中一共包含三类信息：

（1）社交关系信息。包含一个约 256.7 万名微博用户构成的社交网络，其中的社交关系可能是单向的（即单向关注，即为粉丝关系）或双向的（即互相关注，即为好友关系）。

（2）用户微博信息。包含约 4.6 万名用户的微博文本内容，这些用户都属于上述社交网络。

（3）用户标签信息。包含约 0.5 万名用户的年龄、性别及地域标签，均属于上述 4.6 万名用户。

三类信息的关系如图 1 所示：

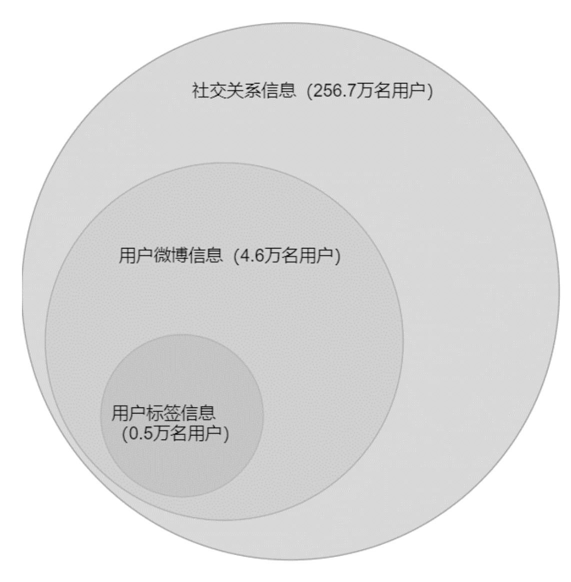


图 1 数据集中三类信息的关系

本文是基于短文本内容的用户性别研究，最终选择了“SMP CUP2016 微博用户画像”比赛数据集中的用户标签信息及其对应的用户微博信息作为本研究的初始数据集，对数据进行预处理工作。

数据预处理分为以下 3 个步骤：

（1）剔除与分析无关的噪声数据。用户微博信息中存在网页链接、字符乱码等噪声数据，这部分数据既不能还原用户的语言表达意图，也不能用于性别预测的特征提取，因此将其剔除。

（2）剔除缺失数据。将缺失性别标签及微博信息少于 5 条的用户标签信息剔除，缺失性别信息的数据无法用于性别预测实验，而微博信息过少也难以提取有效特征，导致性别预测效果差的结果。

（3）对微博信息中的短文本内容进行分词，本研究采用 NLPIR 汉语分词系统进行分词处理，并保留标点符号等原始信息。

经过处理后的数据集包含 4 342 个用户及其发送的微博短文本 331 634 条，用于实验模型的训练与检验。

4 实验构建与特征抽取

4.1 实验构建

本研究的输入为微博短文本，通过对数据进行分析，利用数据特征进行建模，训练相关算法，进而对微博用户的性别进行预测。对性别预测的结果，通过相应评测指标的评价，对算法的效果进行评估。实验的一般流程如图 2 所示：

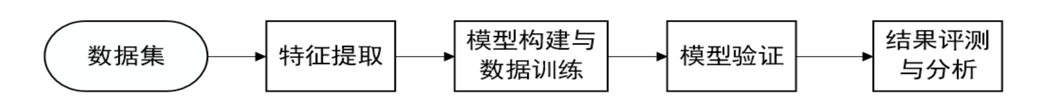


图 2 实验流程

4.2 特征抽取

根据特征抽取方式的不同，可以获得微博短文本内容的两类不同特征，分别为语言特征和主题特征。

4.2.1 语言特征

N. Cheng[28]、D. Rao[34] 在使用 Twitter 数据进行用户性别预测时采纳的语言特征如表 1 所示，考虑到中文文本与 Twitter 用户使用语言的差别，在此基础上，本研究总结了 7 个可从微博短文本中提取的语言特征类别，分别为：①表情：微博中用户使用的表情；②情感词语：积极、消极、焦虑、愤怒等情感词的总称；③语气词：“哈哈”“恩恩”等描述语气的词; ④亲属称呼: “妈妈”“父母”“兄弟姐妹”等称呼；⑤标点符号：包括各种重复使用的标点，如“！！！”; ⑥代词: “你”“你的”等；⑦禁语：指在用户文本中出现的不文明语言。

知识管理论坛，2021（4）:213-227
DOI: 10.13266/j.issn.2095-5472.2021.021

表 1 短文本内容性别预测中使用过的语言特征

D. Rao等[28]	N. Cheng等[34]	D. Bamman等[36]
表情符号	否定词（no,not,never）	代词（you,u,ur）
OMG	积极的情绪（love,nice）	情感词语（sad,love）
省略号	消极的情绪（hurt,ugly）	表情符号（:D）
二元词（my_XXX,）	焦虑（worried,fearful）	亲属称呼（mom,sister）
重复的字母（niceeeeee）	愤怒（hate,kill）	缩写（lol,omg）
自我描述（I_XXX,）	悲伤（crying,grief）	同意（okey,yes）
笑（LOL,ROTFL,haha）	沉思（think,consider）	否定（no,cannot）
愤怒（Ugh,mmmm）	疑惑（maybe,perhaps）	禁语
赞同（yea,yeah,ohya）	肯定（always,never）	介词（a,the,my）
敬语（dude,man,bro,sir）	禁止（block,stop）	
激动（!!!!）	同意（agree,OK,yes）	
单个惊叹号（!）		
困惑（!?!?!）		

本文通过以下方式获取语言特征：

（1）表情。微博短文本中表情以“[具体表情]”的格式体现（例如：[微笑]），可使用正则表达式从文中抓取每一个用户使用的表情，对每一个用户的所有表情取并集获得表情全集。

（2）情感词。对于情感词语语言特征可使用 NTUSD 情感词典与原文进行匹配，获取用户使用的情感词语，对每一个用户的所有情感词语取并集获得情感词语全集。

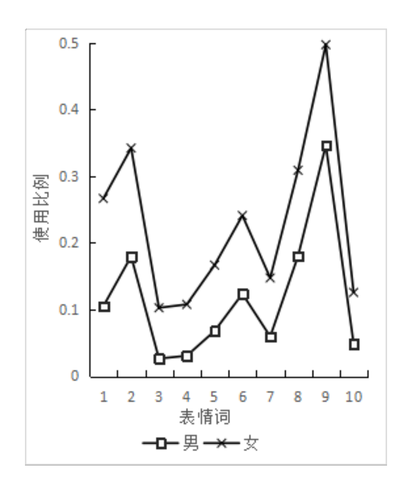
（3）语气词、亲属称呼、标点符号、代词、禁语。由于该类词语的数量相对而言比较少，可以直接通过对部分用户的微博短文本进行标记，找出相关的词语。但考虑到人工标记不全的问题，本文尝试利用文本向量化后的余弦距离，选择相似的词作为该类词语的补充，具体而言：使用 Word2Vec 对分词后的微博短文本进行计算，获得每个词的词向量；针对人工标记出的语气词、亲属称呼、代词、禁语，计算这些词语与语料库中词语的相似度，根据相似度排序筛选出同类别的词作为补充最高的词。

对于短文本中出现的词语 t，使用公式（1）统计性别 i 使用词语 t 的人数占该性别总人数的比例，式中 n（i,t）表示性别 i 的用户中使用了 t 词语的人数，n（i）表示性别 i 的用户总人数。

$$\rho_i(t)=\frac{n(i,t)}{n(i)} \quad i=1,2 \qquad \text{公式（1）}$$

通过对 7 个类别词语在不同性别用户中的使用比例，发现男性和女性使用标点符号和代词类别词语的比例相近，因而不选择这两类词作为语言特征。

对表情、情感词语、语气词、亲属称呼和禁语这 5 个类别的词语分析，男女使用比例最高的 10 个词语的如图 3- 图 7 所示。横坐标代表某个词语，纵坐标为使用比例。从中可以看出：女性相比男性，使用表情的比例更大；情感词语中不同词语的使用情况不同；亲属称呼和语气词中，个别词语男性使用的比例更大，总体上女性更偏向使用该类词语；禁语总体使用比例较小，但男性比女性更倾向使用这类词语。



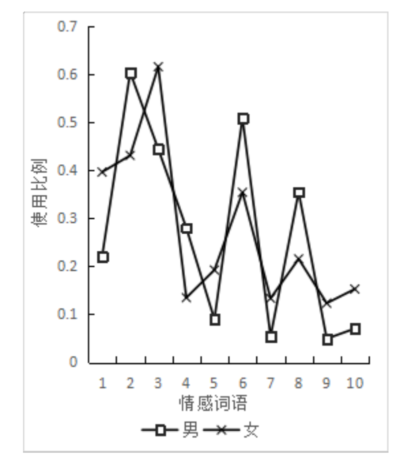
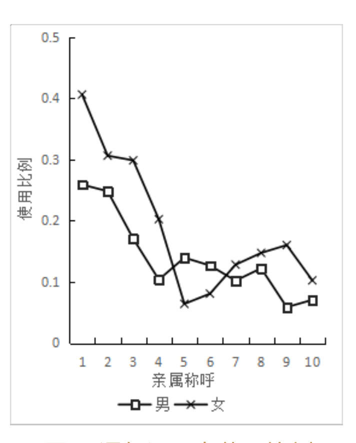
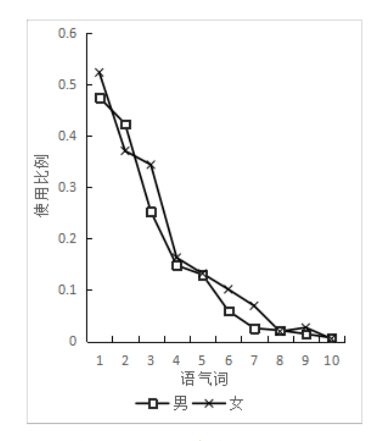


图 3 表情词男女使用比例

图 4 情感词语男女使用比例





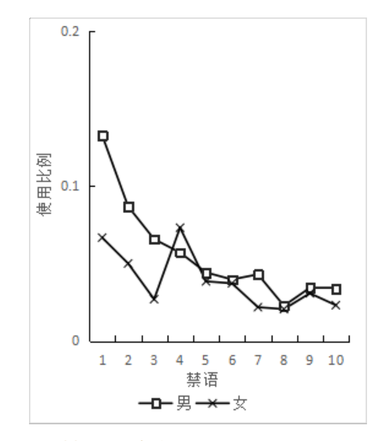


图 5 语气词男女使用比例

图 6 亲属称呼男女使用比例

图 7 禁语男女使用比例

对于表情和情感词语这两个特征，本研究使用卡方检验（chi-square test）进行筛选词语用于特征构建。对词语 t，统计不同性别使用该词语的情况如表 2 所示：

表 2 不同性别使用词语 t 的统计数据

	男性	女性
使用词语t	a	b
未使用词语t	c	d

词语 t 的卡方值 χ^2 可由公式（2）计算得，卡方值越大说明该词语与性别的相关度越大，因此各选择卡方值最大的 100 个词语构成表情和情感词语的语言特征。

$$\chi^2=\frac{n(ad-bc)^2}{(a+c)(b+d)(a+b)(c+d)} \qquad \text{公式(2)}$$

对于语气词、亲属称呼和禁语这三个语言特征，由于在特征词筛选的过程中筛选的词较少，本文不采用上述的卡方检验的方案选取特征，而是将这三个类别的全部词语共计 75 个用于语言特征的构造。

以上 5 个类别共选取了 275 个词语用于构成微博短文本内容的语言特征。对于第 i 个用户，统计该用户使用词语 t 的频次 t_{in}，构建语言特征向量 X_i，其计算公式为：

$$X_i=(t_{i1},t_{i2},t_{i3},\cdots,t_{in}) \qquad \text{公式（3）}$$

4.2.2 主题特征

不同性别用户的兴趣爱好不同会导致发送

知识管理论坛，2021（4）:213-227
DOI: 10.13266/j.issn.2095-5472.2021.021

微博文本的主题不同，因此可以运用LDA（Latent Dirichlet Allocation）模型对用户微博短文本的主题抽取，构建主题特征用于预测用户性别。LDA 是一种基于词袋模型的无监督机器学习方法，可以用来识别大规模文档集中潜藏的主题信息，同时也能有效对文本内容降维，解决数据稀疏问题。

LDA 模型将语料库中的每一篇文档与 K 个主题的多项式分布记为 θ，每个主题与词汇表中的 N 个单词的多项式分布记为 ϕ。θ 和 ϕ 分别有一个带有超参数 α 和 β 的 Dirichlet 先验分布。对于一篇文档 d 中的每一个单词 w_i，$P(z_i=k)$ 代表从文档中抽取一个单词 w_i，$P(w_i|z_i=k)$ 属于主题 z 的概率；从主题 z 中抽取一个单词，代表当取出单词属于主题 k 时该单词为 w_i 的概率。将这个过程重复 N_d 次（N_d 是文档 d 的单词总数），就产生了文档 d。文档中单词 w_i 的概率就能表示为：

$$P(w_i)=\sum_{k=1}^{K}P(w_i|z_i=k)P(z_i=k) \qquad 公式 (4)$$

在本研究中，将每名用户发布的所有短文本内容构成第 i 个用户的文档 D_i，那么可认为文档 D_i 的主题分布向量 $(z_{i1}, z_{i2}, z_{i3}, \cdots, z_{ik})$ 可认为构成了第 i 个用户的主题分布向量。

$$Y_i=(z_{i1}, z_{i2}, z_{i3}, \cdots, z_{ik}) \qquad 公式 (5)$$

本文在 LDA 模型训练的过程中使用困惑度确定模型最佳 K 值，实验过程中，Gibbs 抽样迭代的次数设为 100，α、β 超参数设置为 α=50/K，β=0.01，此时算法有较好的表现[37]。在 K 值提升的过程中，困惑度的下降有限，图 8 展示的是 K 值与困惑度的关系，结合不同 K 值的困惑度和最终产出主题的词语，本文使用 K 值为 15 时产出的模型结果。表 3 展示的是 15 个主题中排序前 10 的词语。

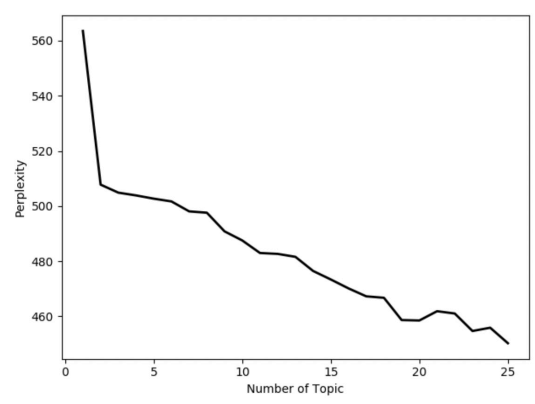


图 8 不同 K 值时困惑度变化

表 3 每个主题的前 10 个词

主题	词语
topic 0	哈哈、嘻嘻、泪、厉害、猫、表情、好看、哥哥、帅、妹妹、...
topic 1	新闻、中国、博文、网易、北京、今日头条、阅读、资讯、专访、微信、...
topic 2	续航、性能、处理器、比亚迪、机型、时速、油耗、变速箱、太阳能、显示器、...
topic 3	礼物、宝贝、情人节、假期、八月、圣诞、节日、晚安、欢迎、看看、...
topic 4	手机、红包、领取、签到、抽奖、小米、信息、iPhone、客户端、相册、...
topic 5	关晓彤、冯小刚、罗志祥、王宝强、陈学冬、邓超、陈赫、李小璐、饰演、孙红雷、...
topic 6	空间、存储、美团、兴趣、公众、微信、水晶、精力、利用、相位、...
topic 7	世界杯、苏宁、奥运会、合同、决赛、冠军、NBA、西班牙、俱乐部、詹姆斯、...
topic 8	技术、学习、效果、训练、能力、专业、个人、项目、挑战、力量、...
topic 9	成都市、河南省、河北省、深圳市、西安市、开发区、广州市、绵阳、市政府、福建省、...
topic 10	美拍、分享、视频、音乐、播放、录制、自、YouTube、魔力、NBA、...
topic 11	智慧、思想、魅力、命运、人才、婚姻、思考、心灵、心态、安全感、...
topic 12	Nike、Party、adidas、运动鞋、配色、Young、Black、Max、Moto、Jordan、...
topic 13	京东、商城、购买、精心、围观、评价、活动、天猫、支付宝、优惠券、...
topic 14	生命、爱情、青春、一生、梦想、时光、人类、过程、内心、意义、...

5 实验结果与分析

5.1 评价方法

研究选用精准率（Precision）、召回率（Recall）和 F 值（F-Measure）作为评价指标来对实验的结果进行比较评价。三种指标的计算方式如下：

$$P = \frac{TP}{TP+FP}; \qquad \text{公式（6）}$$

$$R = \frac{TP}{TP+FN}; \qquad \text{公式（7）}$$

$$F-Measure = \frac{2*P*R}{P+R} \qquad \text{公式（8）}$$

以女性性别为例，TP 表示将性别预测正确的数量；FN 表示将正确的女性预测为男性的数量；FP 表示将正确的男性预测为女性的数量。

5.2 模型训练

5.2.1 训练数据与测试数据

数据预处理得到的 4 342 名用户中男性和女性的数据比例不一致，为更好地进行试验，随机选择 2 110 名用户按照 1：1 的性别比例构建实验数据集，2 110 名用户共发表微博 156 627 篇。其中 1 560 名用户用于模型的训练（男女性别比例为 1：1），550 名用户用于模型效果的检验（男女性别比例为 1：1）。

在模型训练阶段，1 560 名用户采用 5 折交叉检验的方法进行模型训练，保证数据的充分利用与模型训练的准确。

5.2.2 模型的参数调优

将抽取的用户语言特征与主题特征组合成为新的特征向量进行实验，获取最佳的性别预测结果。

$$M_i=(X_i+Y_i)=(t_{i1}, t_{i2}, t_{i3}, \cdots, t_{in}, z_{i1}, z_{i2}, z_{i3}, \cdots, z_{ik}) \qquad \text{公式（9）}$$

本研究采用的是支持向量机（Support Vector Machine, SVM）这一基于统计学习理论的机器学习方法。支持向量机通过核函数解决计算复杂度的问题，除重要的参数 cost 外，还有四种不同的核函数，分别为线性（Linear）核函数、径向基（radial basis function, RBF）核函数、sigmoid 核函数和多项式（Polynomial）核函数，每一种核函数有不同数量的参数。本文使用 LIBSVM 这一软件包实现对用户性别的预测，通过选定不同的核函数、控制相关变量对核函数进行参数训练，从而获得最优的预测效果。

对于线性核函数只需训练参数 cost。为了使 cost 值尽量覆盖更多的值，本文使用指数函数规定 cost 的选取范围，其取值范围为 2^{-10} 至 2^5。最终结果显示当 cost=1/32 时在评价指标上表现最好，有较好的预测效果。图 9 展示了不同 cost 取值时的预测效果，可以看出当 cost 值较小或者较大的时候，预测的效果都不够好，这是因为，cost 值越高越容易过拟合，cost 值越小越容易欠拟合。

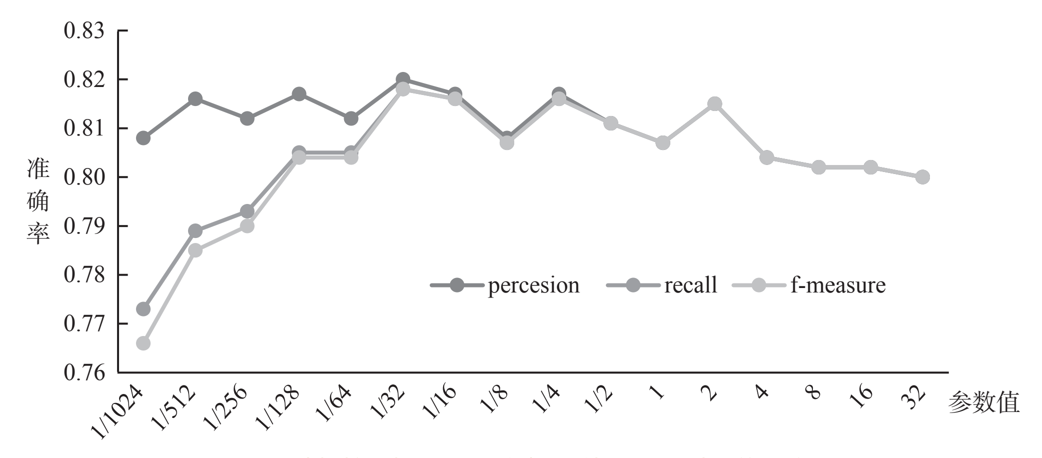


图 9 线性核函数 cost 取值变化对性别预测结果的影响

知识管理论坛，2021（4）:213-227
DOI: 10.13266/j.issn.2095-5472.2021.021

径向基核函数有 gamma 参数以及 cost 参数，本研究使用 GridSearch 网格搜索的方式确定最佳参数，gamma 以及 cost 的变化范围都是从 2^{-10} 至 2^{5}。当 cost=32，gamma=1/128 时预测结果最佳。gamma 是 RBF 函数中自带的一个参数，一定程度上决定了数据映射到新的特征空间后的分布，gamma 值越大支持向量越少，gamma 值越小支持向量越多，支持向量的个数影响模型训练的速度和准确度。图 10 展示的是固定 cost 值为 1，改变模型 gamma 的值，在测试集中进行分类的结果，从中可以看到，当 gamma 大于 1 的时候预测的准确率很低。

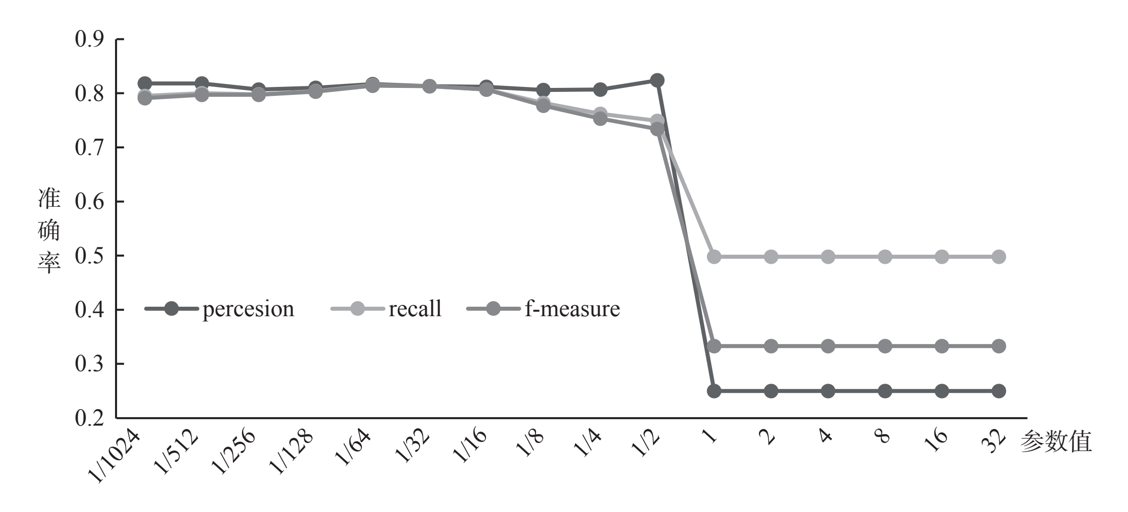


图 10 RBF 核函数 gamma 值变化对预测结果的影响

sigmoid 核函数有 cost、gamma 和 coef0 三种参数，本研究分两步进行参数调优：①将 cost 设为默认值 1，使用 GridSearch 网格搜索确定 gamma 以及 coef0 的值，其中 gamma 和 coef0 的取值范围定为 2^{-10} 至 2^{5}；②使用第一步训练出的 gamma 以及 coef0 值，将 cost 的范围设定为 2^{-10} 至 2^{5} 进行训练。最终得到当 cost=32，coef0=8，gamma=1/16 时模型的预测效果最佳。图 11 展示的是固定 cost 值与 gamma 值，改变模型中 coef0 的值对测试集的预测效果，当 coef0 的值超过某个值后，其预测效果将大幅下滑，通常情况下 coef0 的值越大，预测结果越差。

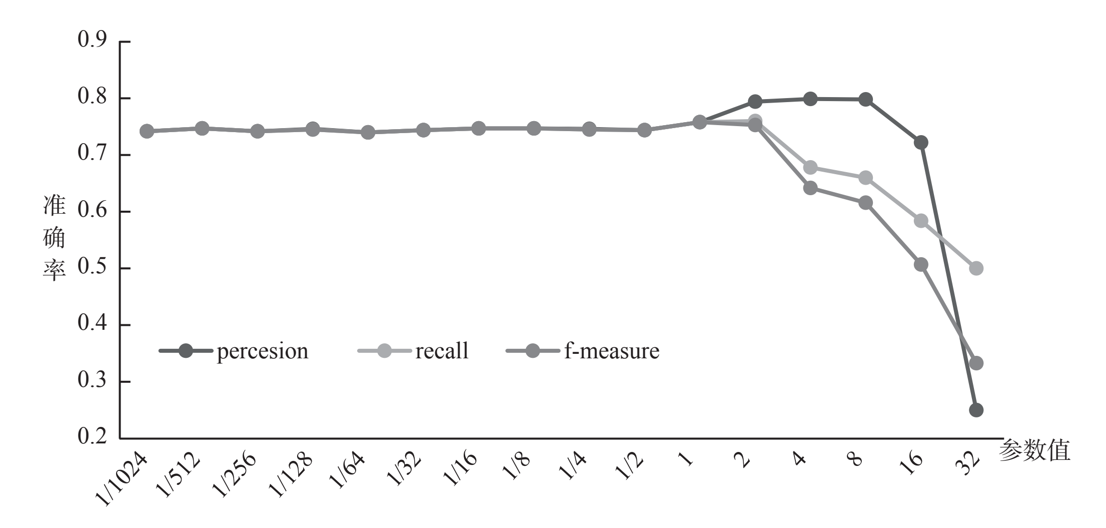


图 11 sigmoid 核函数 coef0 变化对预测结果的影响

多项式核函数有 cost、gamma、coef0 和 degree 4 种参数，其中 degree 参数最为关键。本文分 3 个步骤来确定最佳参数：①将 cost，gamma，coef0 设定成为默认值，将 degree 范围设定为 0 至 19 进行训练，得到最佳 degree 值为 1；②将 cost 设置为默认值，degree 设置为最优参数 1，使用 GridSearch 网格搜索法使 gamma 及 coef0 在 2^{-10} 至 2^5 取值范围内变化，得到最佳的 gamma=1/4, coef0=16 的值；③ degree=1，gamma=1/4,coef0=16 设为固定参数，将 cost 取值在 2^{-10} 至 2^5 训练，最终确定的最优参数为 degree=1， gamma=1/4，coef0=16,cost=16 时模型的预测效果最佳。图 12 展示的是改变模型中 degree 的值，对测试集进行预测的效果，其中 degree 的变化范围从 0 到 19，随着 degree 值越来越大，预测效果越来越差，当 degree 超过 15 后预测结果几乎没有任何改变。

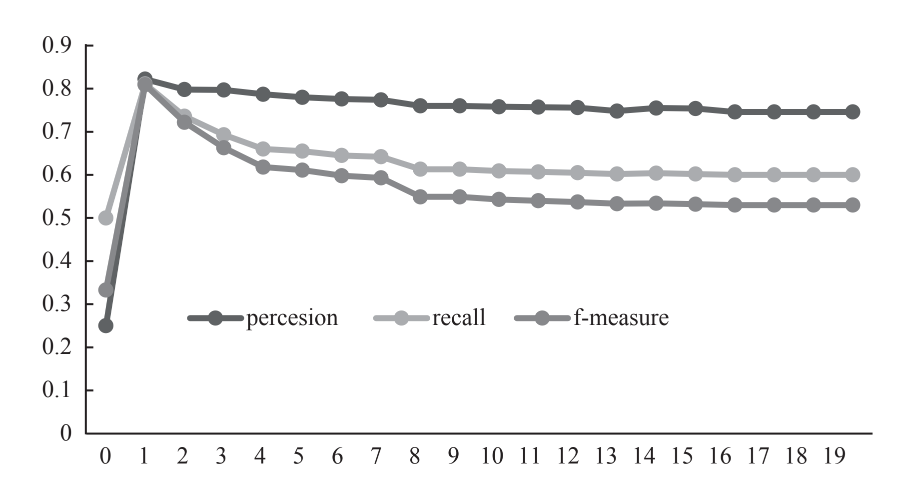


图 12 多项式核函数 degree 值变化对预测效果的影响

针对在测试集的预测结果，选取 4 种不同核函数效果最优的参数进行横向比较，可以看出 sigmoid 核函数的表现最差，在三个指标中均未达到 80%；径向基核函数的预测效果最好，在三个评测指标中都比其他核函数表现更好。因此将选择参数为 cost=32，gamma=1/128 的径向基核函数作为预测模型，用于实验数据的预测。

表 4 4 种核函数的最优参数及预测效果对比

评测指标	线性核函数（cost=1/32）	径向基核函数（cost=32，gamma=1/128）	sigmoid 核函数 (cost=32,coef0=8,gamma=1/16)	多项式核函数 (degree=1，gamma=1/4,coef0=16,cost=16)
精准率	0.82	**0.829**	0.791	0.822
召回率	0.818	**0.829**	0.787	0.811
F 值	0.818	**0.829**	0.787	0.809

5.3 结果比较

5.3.1 baseline 选择

基于 n 元语法模型的性别预测方法 [31-34] 和基于心理语言学词典的性别预测方法 [38] 都是利用用户的微博文本内容进行性别预测的自然语言处理方法，在针对社交媒体中用户的性别预测有较好的效果。本文选择这两种方法作为 baseline 进行比较。

知识管理论坛，2021（4）:213-227
DOI: 10.13266/j.issn.2095-5472.2021.021

在 n 元语法模型中，通过抽取 500 个最具有区分性的一元和二元词，统计每名用户的使用频率作为权重构建用户的特征向量；针对基于心理语言学词典的用户特征向量，使用文心（TextMind）中文心理分析系统 [39] 构建，对用户发文的内容进行统计，提取 102 个特征，包括各种词性词语使用的数量、词长比例、情感词数量等。

5.3.2 结果对比

将实验数据应用于训练所得的最优模型，如图 13 所示，本文提出的主题特征、语言特征构建及两种特征融合构建的性别预测模型的精准率、召回率和 F 值指标均比选择的 baseline 有所提升，特别是与心理语言学词典相比，提升较大，本研究表现最差的主题特征在该指标上都提升了 14.3 个百分点。

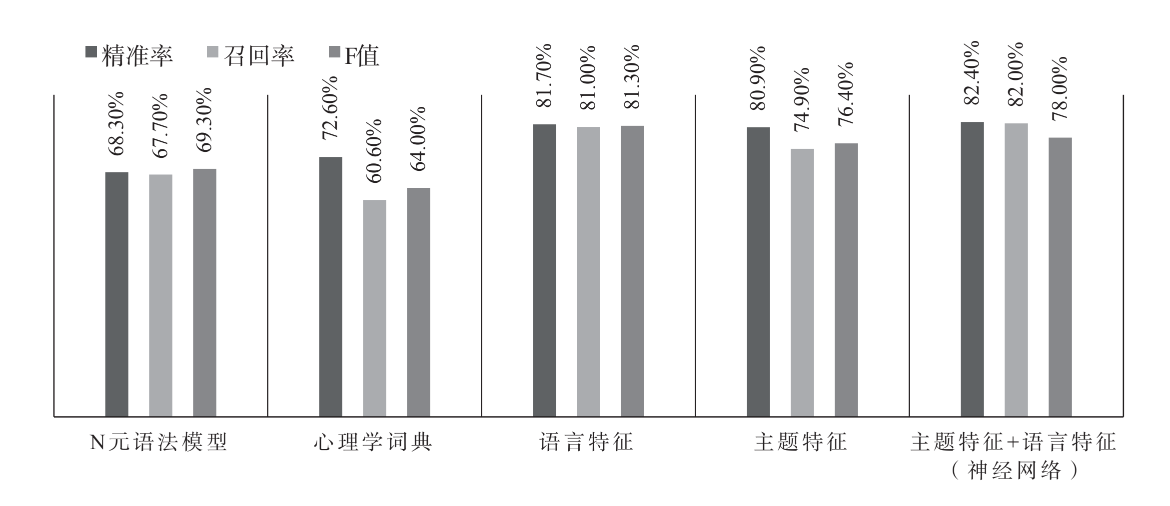


图 13 不同特征的实验效果比较

基于 n 元语法模型的性别预测效果不显著，精准率、召回率和 F 值都未达到 70%，其中 F 值表现最好，为 69.3%。通过分析可知，n 元语法模型虽然抽取了 500 个特征进行特征向量的降维，但构造的特征向量依然较为稀疏。表 5 展示了针对同一用户使用 n 元语法模型和语言特征构造的向量。由于 n 元语法模型是针对所有的一元和二元词汇进行的特征选择，这些词语数量较多，造成向量稀疏。而本文构建语言特征时选择的词语，通过对用户使用频率的统计有效避免了稀疏问题。

表 5 使用 n 元语法特征与语言特征构造的向量对比

使用n元语法模型构造的特征向量	使用语言特征构造的特征向量
0,0,1,1,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,...	0,0,0,7,2,0,0,1,0,0,0,0,11,0,0,0,3,1,2,1,0,0,3,...

心理语言学词典方法的精准率虽然达到了 72.6%，但召回率只有 60%。通过分析可知，心理语言词典构建的特征中包含代词、表达符合这类的词语，而本文的语言特征通过统计这类词语与性别的关联度，这类词语忽略，不纳入语言特征的构建，而心理语言词典没有忽略，均纳入了特征构建，得到的精确率、召回率和 F 值比语言特征分别低 9.1%、20.4% 和 17.3%。从而进一步验证了基于语言特征构建模型预测性别时需忽略代词和表达符合等。

对比本研究的主题特征、语言特征和两种特征叠加可知，主题特征表现最差，语言特征表现较好，叠加特征结果最优。在精准率指标上，语言特征的精准率为 81.7%，仅比主题特征高 0.8%，但在召回率和 F 值上，语言特征大幅提升，分别提升了 6.1% 和 5.7%。精准率的提升，

表明语言特征进行性别预测时更加有效。两种特征叠加的预测结果，在语言特征的基础上精准率进一步提升了 1.4%，达到 83.1% 提升效果显著；相较之下，召回率和 F 值与语言特征相比提升有限。分析可知这与主题特征的特征数量与预测效果有关，一方面主题特征的特征数量较少，另一方面主题特征的召回率与 F 值相对语言特征差值较大，因此两种特征叠加对召回率和 F 值的提升较少。

同时，本文对比了 SVM 模型与 BP 神经网络和 TEXTCNN[40] 神经网络的效果。本文构建了 2 层隐藏层的 BP 神经网络：第一层含有神经元 120 个，第二层有神经元 60 个，使用通过主题特征和语言特征提取的向量作为输入，使用 sigmoid 函数作为输出层函数。对于 TEXTCNN 模型，则不再使用特征向量作为输入，而是用户发送的文本分词后的词向量，向量的维数为 128 维；在卷积层，使用三种不同高度的卷积核，分别为 2、3、4，每一种卷积核的个数设置为 128 个。两种不同模型与 SVM 模型的效果对比如图 14 所示，总体而言三种模型的效果较为接近，SVM 的效果最好。SVM 模型的 F 值比神经网络高了 4%，比 TEXTCNN 模型高了 2%，精准率上 SVM 模型比 BP 神经网络和 TEXTCNN 高 1%。TEXTCNN 的效果比较优秀也是因为模型考虑到了语言上下文之间的关系，而通过语言特征和主题特征提取的向量也有相同效果，进一步说明了语言、主题两类特征对于文本性别分类的重要性。

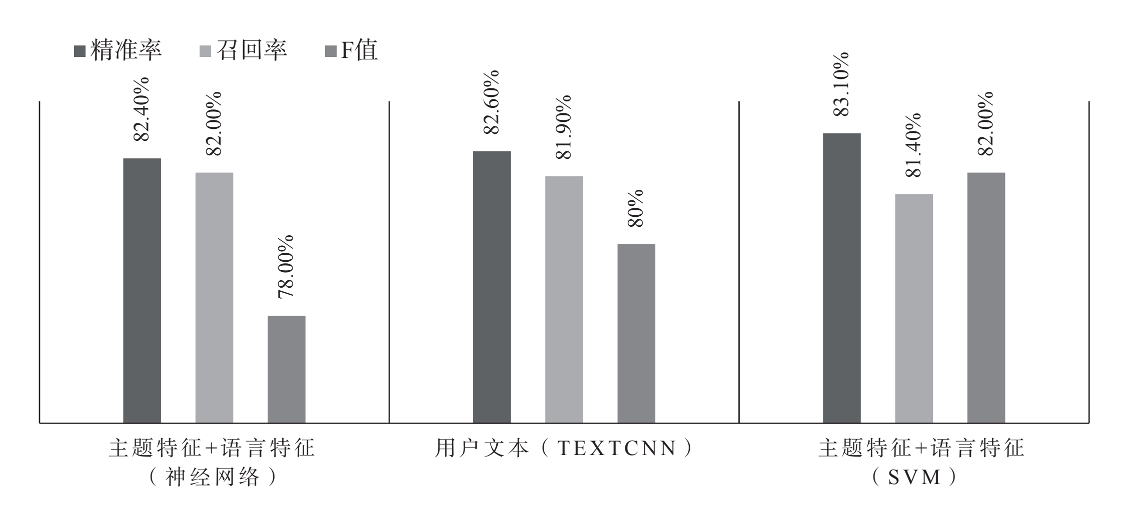


图 14 神经网络的对比

总体而言，本文提出的主题特征、语言特征和两种特征叠加对性别的预测均优于选取的 baseline 方法，对社交媒体用户性别的预测效果起到了很好的提升。

6 结语

社交媒体中个人信息的隐私保护始终面临诸多挑战，虽然已有法律条文的规范，但在实践过程中用户依然暴露在风险中。利用社交媒体中的相关信息进行性别预测，能对用户起到一定的保护作用。

本文以中文社交媒体新浪微博为例，从用户的短文本中提取主题特征和语言特征，对支持向量机的机器学习算法进行参数调优与训练，得到一个对性别预测有显著提升的分类器，起到了较好的预测效果，在精准率、召回率和 F 值上都有所提升，特别是精准率与 baseline 方法相比提升均超过 10 个百分点，说明从短文本的角度对用户性别进行预测是一个有效的途径。同时，与常用的 n 元语法模型和心理语言学词典方法相比较，有效解决了构造向量的稀疏问题，为进一步促进基于性别的用户信息保护提

ChinaXiv合作期刊
知识管理论坛，2021（4）:213-227
DOI：10.13266/j.issn.2095-5472.2021.021

供了基础。

本研究提出的方法是利用中文短文本进行性别预测，该方法可推广到其他社交媒体如 Twitter 中进行中文用户的性别预测。

参考文献：

[1] 陈传夫 , 刘雅琦 . 公共部门信息增值利用中的个人信息保护 [J]. 情报科学 , 2010, 28(10): 1455-1460.

[2] 刘雅琦 . 公共部门信息增值利用中的个人信息保护立法研究 [J]. 情报理论与实践 , 2011, 34(4): 40-43.

[3] 郑莉 , 蔡琼 , 石曼 , 等 . 社交网络隐私成本的量化研究 [J]. 科教导刊 (电子版), 2019(1): 282.

[4] 曹杨 . 微博用户性别分类研究及应用 [D]. 合肥 : 安徽大学 , 2019.

[5] 熊杰 . 政务微博在线评论中的用户情绪及行为研究 [D]. 成都 : 电子科技大学 , 2020.

[6] WALTON S C, RICE R E. Mediated disclosure on Twitter: the roles of gender and identity in boundary impermeability, valence, disclosure, and stage[J]. Computers in human behavior, 2013, 29(4): 1465-1474.

[7] PIAO G, BRESLIN J G. User modeling on Twitter with WordNet Synsets and DBpedia Concepts for Personalized Recommendations[C]//ACM international conference on information & knowledge management. Indianapolis: ACM, 2016：2057-2060.

[8] PAN. Shared tasks[EB/OL].[2021-02-04]. https://pan.webis.de/shared-tasks.html.

[9] BIENDATA. 比 赛 项 目 [EB/OL].[2021-02-04]. https://www.biendata.xyz/competition/.

[10] SMITH J. Gender prediction in social media[EB/OL].[2021-02-04]. https://arxiv.org/abs/1407.2147.

[11] ABBASI M A, CHAI S K, LIU H, et al. Real-world behavior analysis through a social media lens[C]//International conference on social computing, behavioral-cultural modeling, and prediction. Berlin: Springer, 2012: 18-26.

[12] ZHELEVA E, GETOOR L. To join or not to join: the illusion of privacy in social networks with mixed public and private user profiles[C]//Proceedings of the 18th international conference on World Wide Web, 2009: 531-540.

[13] SCHWARTZ H A, EICHSTAEDT J C, KERN M L, et al. Personality, gender, and age in the language of social media: the open-vocabulary approach[J]. PloS one, 2013, 8(9): e73791.

[14] VICENTE M, BATISTA F, CARVALHO J P. Gender detection of Twitter users based on multiple information sources[M]//Interactions between computational intelligence and mathematics part 2. Cham: Springer, 2019: 39-54.

[15] SUN X, WU P, LIU H. Facial age estimation using bio-inspired features and cost-sensitive ordinal hyperplane rank[C]// IEEE, International Conference on Cloud Computing and Intelligence Systems. Shenzhen: IEEE, 2015:81-85.

[16] GUO G, MU G, FU Y. Gender from body: a biologically-inspired approach with manifold learning[M]// Computer vision – ACCV 2009. Berlin: Springer, 2009.

[17] LANITIS A, TAYLOR C J, COOTES T F. Toward automatic simulation of aging effects on face images[J]. Pattern analysis & machine intelligence IEEE transactions on, 2002, 24(4):442-455.

[18] GUNAY A, NABIYEV V V. Automatic age classification with LBP[C]// International symposium on computer and information sciences. Istanbul: IEEE, 2008:1-4.

[19] SHAN C. Learning local binary patterns for gender classification on real-world face images[M]. Amsterdam: Elsevier Science Inc. 2012.

[20] BALUJA S, ROWLEY H. Boosting sex identification performance[J]. International journal of computer vision, 2007, 71(1): 111-119.

[21] MANSANET J, ALBIOL A, PAREDES R. Local deep neural networks for gender recognition[M]. Amsterdam: Elsevier Science Inc, 2016.

[22] 吴泽银 . 基于集成卷积神经网络的人脸性别识别研究 [D]. 广州 : 华南理工大学 ,2016.

[23] BURGER J D, HENDERSON J, KIM G, et al. Discriminating gender on Twitter[C]// Conference on empirical methods in natural language processing. Edinburgh: Association for Computational Linguistics, 2011: 1301-1309.

[24] ALOWIBDI J S, BUY U A, YU P. Language independent gender classification on Twitter[C]// IEEE/ACM international conference on advances in social networks analysis and mining. Niagara Falls: IEEE, 2013:739-743.

[25] 钱铁云 , 尤珍妮 , 陈丽 , 等 . 基于兴趣标签的缄默用户性别预测研究 [J]. 华中科技大学学报 (自然科学版), 2015, 43(12): 101-105.

[26] LI S, WANG J, ZHOU G, et al. Interactive gender inference with integer linear programming[C]//

International joint conference on artificial intelligence. Barcelona: AAAI Press, 2015: 2341-2347.

[27] 戴斌 , 李寿山 , 贡正仙 , 等 . 基于多类型文本的半监督性别分类方法研究 [J]. 山西大学学报 (自然科学版), 2017, 40(1):14-20.

[28] CHENG N, CHANDRAMOULI R, SUBBALAKSHMI K P. Author gender identification from text[J]. Digital investigation, 2012, 8(1):78-88.

[29] FILHO J A B L, PASTI R, CASTRO L N D. Gender classification of twitter data based on textual meta-attributes extraction[C]// World conference on information systems and technologies. Switzerland: Springer, 2016:1025-1034.

[30] WANG Q, MA S, ZHANG C. Predicting users' demographic characteristics in a Chinese social media network[J]. The electronic library, 2017, 35(4): 758-769.

[31] PEERSMAN C, DAELEMANS W, VAERENBERGH L V. Predicting age and gender in online social networks[C]// International CIKM workshop on search and mining user-generated contents. Glasgow：DBLP, 2011:37-44.

[32] 王晶晶 , 李寿山 , 黄磊 . 中文微博用户性别分类方法研究 [J]. 中文信息学报 , 2014, 28(6):150-155.

[33] MILLER Z, DICKINSON B, HU W. Gender prediction on Twitter using stream algorithms with N-Gram character features[J]. International journal of intelligence science, 2012, 2(4):143-148.

[34] RAO D, YAROWSKY D, SHREEVATS A, et al. Classifying latent user attributes in Twitter[C]// International workshop on search and mining user-generated contents. New York: ACM, 2010:37-44.

[35] BIENDATA.SMPCUP2016 微博用户画像数据 [EB/OL]. [2020-10-08].https://www.biendata.xyz/competition/smpcup2016/data/.

[36] BAMMAN D, EISENSTEIN J, SCHNOEBELEN T. Gender identity and lexical variation in social media[J]. Journal of sociolinguistics, 2014, 18(2):135–160.

[37] BLEI D M, NG A Y, JORDAN M I. Latent Dirichlet allocation[J]. Journal of machine learning research, 2003, 3(3):993-1022.

[38] CHEN J, HUANG H, TIAN S, et al. Feature selection for text classification with Naïve Bayes[J]. Expert systems with applications an international journal, 2009, 36(3):5432-5435.

[39] GAO R, HAO B, LI H, et al. Developing simplified Chinese psychological linguistic analysis dictionary for Microblog[M]// Brain and health informatics, 2013:359-368.

[40] KIM Y. Convolutional neural networks for sentence classification[EB/OL].[2021-02-04]. https://arxiv.org/abs/1408.5882

作者贡献说明：

雅琦：实验设计与论文修改;

李得志：数据收集、实验与部分论文撰写;

王瑞雪：数据分析与部分论文撰写。

ChinaXiv合作期刊
知识管理论坛，2021（4）:213-227
DOI: 10.13266/j.issn.2095-5472.2021.021

Research on Gender Prediction of Chinese Social Media Users
——Taking Sina Weibo Short Text Content as an Example

Liu Yaqi[1] Li Dezhi[2] Wang Ruixue[3]
1. School of Information and Security Engineering, Zhongnan University of Economics and Law, Wuhan 430073
2. Baidu Network Technology Co.,Ltd., Beijing 100085
3. School of Information Management, Wuhan University, Wuhan 430072

Abstract: [Purpose/significance] Different from the rapid development of the Internet, the development of personal information security protection is relatively lagging. By predicting the gender of social media users, it can better provide privacy protection for the users. **[Method/process]** The short texts posted by users in social media, Sina Weibo, were taken as the research object. The experiment extracted linguistic features and topic features from the short texts. For each user, we constructed features vector based on linguistic features, topic features, and the superposition of two features, then used SVM Machine learning algorithms built a classifier for gender prediction. **[Result/conclusion]** Experiments show that the linguistic features and topic features can predict the gender of the users accurately, and the effect is superior to other features used in gender prediction.

Keywords: short text gender prediction topic features linguistic features